The Journal of Open Source Software

SCOPRO: an R package for assigning score projection between query and reference from single-cell datasets

Gabriele Lubatti[*1,2,3] **and Antonio Scialdone**[1,2,3¶]

1 Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, Munich, Germany **2** Institute of Functional Epigenetics, Helmholtz Zentrum München, Neuherberg, Germany **3** Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

In the last decade the size of datasets generated with single cell RNA sequencing technique has grown exponentially. Large references, like the human cell atlas, with annotated cell types, are nowadays available (Regev A, 2017). Efficient tools are essential in order to characterize a cell c from a newly generated dataset in quick and robust way, by projecting the new (query) dataset into an existing reference dataset. Several tools were developed in the last years to address this task (Kiselev, n.d.; Li, 2020; Stuart T, 2019). However, while they perform well if the similar cell types are present in both reference and query, they tend to fail if a cell type is only in the query but not in the reference. The main reason is that these methods predict always a label for the query cell, even if it is from cell type not included in the reference. Moreover the features (genes) that are in common between the reference cluster and the query cell that lead to the labelling are not given as output. Here we propose SCOPRO, an R library that assigns an absolute score (from 0 to 1) between each cluster in the query dataset and a given cluster in the reference dataset. The score is given by the fraction of genes that are conserved between the query and the reference cluster. Since is bounded between 0 and 1, the score is comparable across clusters and does not depend on the clusters included in the reference dataset.

Statement of need

As more and more new datasets are generated with single cell RNA sequencing technique, it has become crucial to compare the new datasets with already existing and annotated references. In the last years, several tools were developed to perform label transfer from a reference to a query dataset. Among the most popular ones there are Seurat, SciBet and scmap [Stuart T (2019); Li (2020); Kiselev (n.d.);]. They are all implemented in an R package. Seurat(Stuart T, 2019) is based on the idea of using anchors between query and reference. Anchor is a cells pair (one from query, one from reference) made up of mtual nearest neighbors(Haghverdi, 2018) found in a shared low dimensional embedding. Once the anchors are identified, the annotation of each cell in the query set is achieved using a weighted vote classifier based on the reference cell identities. So for each query cell a quantitative score for every cluster in the reference dataset is given. In SciBet(Li, 2020), first a features selection process is done for each cell types in the reference with E-test. Then a multinomial model is built (one for each cell type in the reference). The parameters of the distribution are computed starting from the normalized expression of the selected features. The query cell is annotated with the cell type in the reference that maximized the likelihood function. Scmap(Kiselev, n.d.) identifies for each query cell the closest cluster in the reference (represented with a centroid given by a vector of the median value of the expression of each gene) with nearest neighbor approach. The Similarities

*first author

between the query cell and the closest reference cluster are computed using cosine similarity and Pearson and Spearman correlations. If at least two of the similarities are in agreement, and if at least one is above 0.7, then the query cell is labelled as the closest reference cluster. Otherwise the cell is labelled unassigned. Seurat and scmap assign a quantitative score to each of the query cell. However this quantitative score depends on the clusters included in the reference dataset. On the contrary SciBet returns as output only the predicted label of the query cell, but not a quantitative score. For all the three methods the labelling show simply to which cluster in the reference the query cell is closer to, but not how much the query cell is similar to the reference. Another limitation of the previous methods is that the common genes that drive the labelling of the query cell are not given as output. To overcome these limitations, we develop SCOPRO, an R package that assigns a score projection from 0 to 1 between a given cluster in the reference and each single cluster from a query dataset. The score is assigned based on the fraction of specific markers of the reference cluster that are conserved in the query cluster. The first step is to select as features only the markers of the reference clusters with a median above a given threshold in one cluster and below this threshold in all the other clusters. For a given cluster, a connectivity matrix is computed with number of rows and number of columns equal to the number of the selected markers. Each entry (i,j) in the matrix can be 1 if the fold change between gene i and gene j is above a given fold change. Otherwise is 0. Finally the connectivity matrix of the reference cluster and all the clusters in the query dataset are compared. A gene i is considered to be conserved between a reference cluster and a query cluster if the jaccard index of the links of gene i is above a given threshold. SCOPRO returns as output a score between 0 and 1 that rely only on the fraction of conserved genes between the reference and the query clusters, but not on which clusters are included in the reference. For this reason, the score from SCOPRO can be interpreted as an attempt to provide an absolute measure of similarity between query and reference clusters, differently from the output of the previous methods. In addition SCOPRO provides as output the genes that are conserved between query and reference dataset. These genes are relevant because they are responsible or the final score.

Key functions

The two main functions of SCOPRO are: 1. 'SCOPRO': It takes as input the normalized count matrix of the query and reference datasets, the unsupervised cluster assignment for the query, the selected reference cluster for which we want to compute the score and the features (markers genes) from the reference. The output returns by SCOPRO is a list including the score and the conserved genes between each query cluster and the given reference 2. 'cluster.plot_score': It takes as input the output of SCOPRO and it returns the score between each query clusters and the reference. 3. 'plot_score_genes': it returns a balloon plot with the conserved and not conserved genes between a given reference cluster and the query clusters.

In SCOPRO package are also implemented wrapper functions for popular R based projection tools (Seurat, SciBet and scmap). These wrapper functions are built in order that their output is perfectly integrable with other SCOPRO functions This has the advantage of having in one, easy to use library several methods that can be used for a comparison with the output of SCOPRO. As example to show how SCOPRO works we used as reference a dataset from mouse embryo development (in vivo dataset)(Deng et al., 2014; Mohammed et al., 2017) including stage 2-cells stages and epiblast stages (from 4.5 to 6.5). As query we used a mouse embryonic stem cells dataset (in vitro dataset)(Iturbide, 2021). We run SCOPRO selecting as reference cluster the late 2-cells stage. We noticed that the query cluster 2 has a very high score for late 2-cells stage, while the score for cluster 0 and 1 is very low (**figure 1**). Interestingly among the conserved markers between cluster 2 and the late 2-cells stage there are Zscan4 family genes (**figure 2**). It is known that in mouse embryonic stem cells, a rare population of cells with typical markers of late 2-cells stage including Zscan4 genes is present. This population is called 2 cells like cells (2CLC)(Macfarlan TS, 2012; Rodriguez-Terrones, 2018). Therefore cluster 2 in our query dataset is the 2CLC.

The Journal of Open Source Software

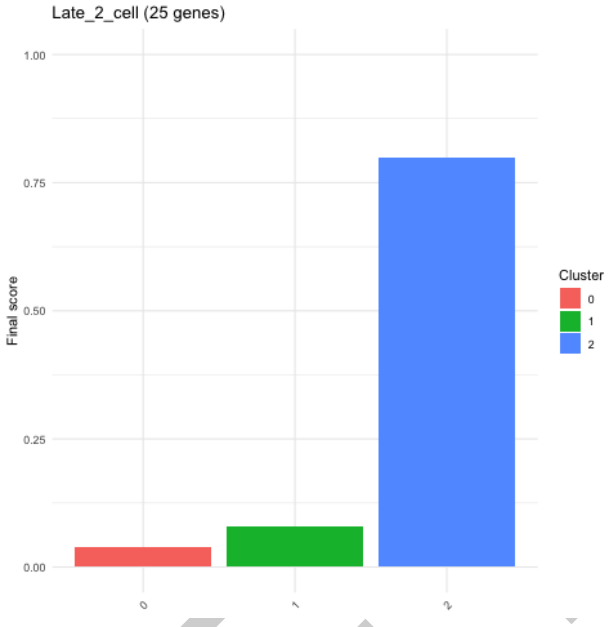

Figure 1: Score given by SCOPRO for clusters from mouse embryonic stem cells from Iturbide (2021). The reference cluster is late 2 cells stage from mouse embryo development from Deng et al. (2014); Mohammed et al. (2017).

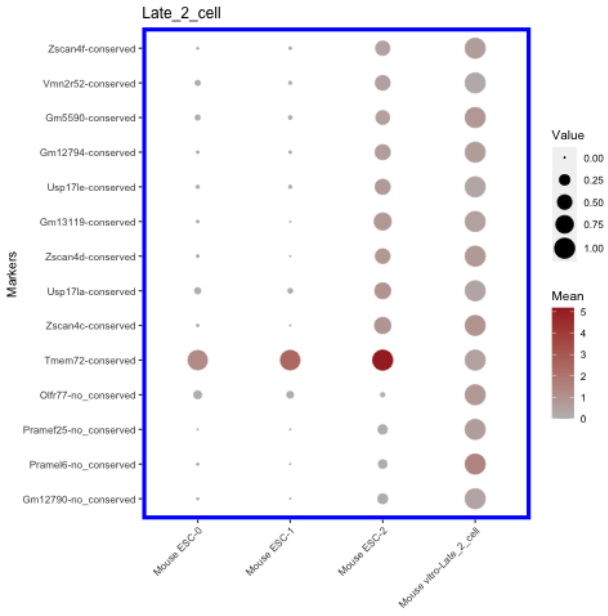

Figure 2: Balloon plot with the genes conserved and not conserved between cluster 2 from mouse embryonic stem cells and the late 2 cells stage.

The advantage of SCOPRO in comparison with other published methods is that it assigns a score that does not depend on the clusters present in the refence dataset. Starting from the same query, we run again SCOPRO and Seurat but this time removing from the reference dataset the cluster late 2 cells stage. If only epiblast stages from 4.5 to 6.5 are used, then Seurat will still assign cluster 2 to epiblast 4.5 and epiblast 5.5 (**figure 3**, **figure 4**), although this cluster shares just a few markers with these reference stages. On the other hand, SCOPRO

correctly assigns a low score (below 0.5) in cluster 2 for both epiblast 4.5 and epiblast 5.5 (**figure 5**).

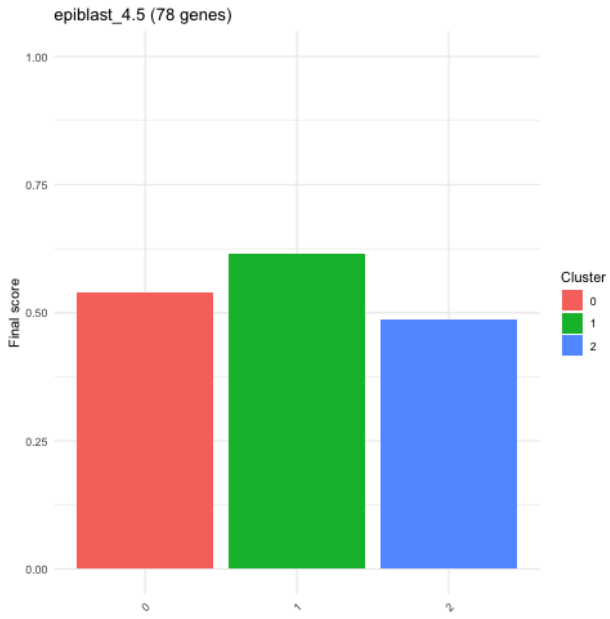

Figure 3: Score given by SCOPRO for clusters from mouse embryonic stem cells from Iturbide (2021). The reference cluster is epiblast 4.5 stage from mouse embryo development from Deng et al. (2014); Mohammed et al. (2017).

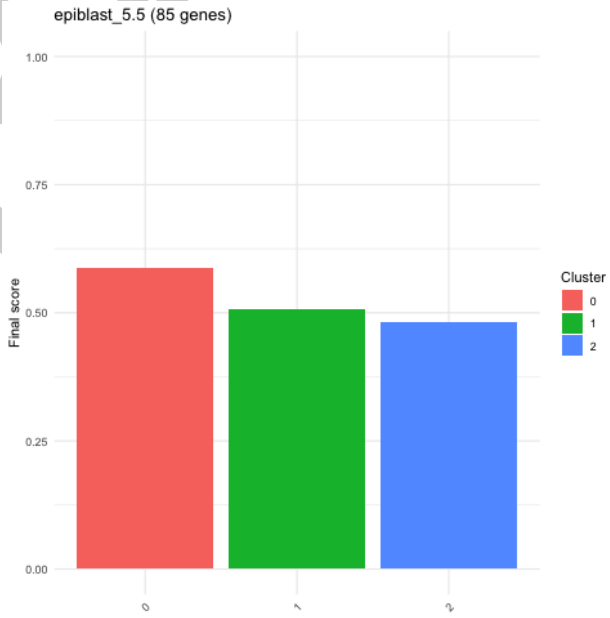

Figure 4: Score given by SCOPRO for clusters from mouse embryonic stem cells from Iturbide (2021). The reference cluster is epiblast 5.5 stage from mouse embryo development from Deng et al. (2014); Mohammed et al. (2017).

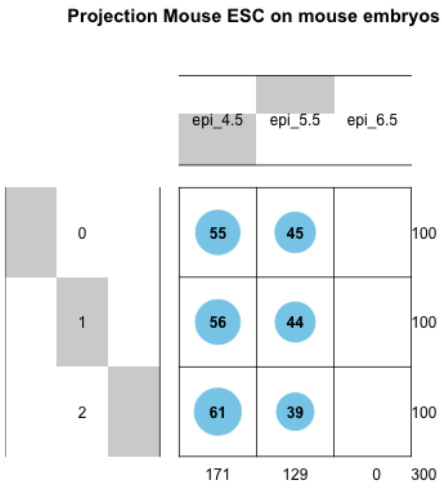

Figure 5: Score given by Seurat for clusters from mouse embryonic stem cells from Iturbide (2021). The reference dataset include epiblast 4.5, epiblast 5.5 and epiblast 6.5 from mouse embryo development from Deng et al. (2014); Mohammed et al. (2017).

References

Deng, Q., Ramsköld, D., Reinius, B., & Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, *343*(6167), 193–196. https://doi.org/10.1126/science.1245316

Haghverdi, L., L. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*. https://doi.org/10.1038/nbt.4091

Iturbide, R. T. S., A. (2021). Retinoic acid signaling is critical during the totipotency window in early mammalian development. *Nature Structural & Molecular Biology*. https://doi.org/10.1038/s41594-021-00590-w

Kiselev, Y., V. (n.d.). Scmap: Projection of single-cell RNA-seq data across data sets. *Nature Methods*. https://doi.org/10.1038/nmeth.4644

Li, L., C. (2020). SciBet as a portable and fast single cell type identifier. *Nature Communications*. https://doi.org/10.1038/s41467-020-15523-2

Macfarlan TS, D. S., Gifford WD. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*. https://doi.org/10.1038/nature11244

Mohammed, H., Hernando-Herraez, I., Savino, A., Scialdone, A., Macaulay, I., Mulas, C., Chandra, T., Voet, T., Dean, W., Nichols, J., Marioni, J. C., & Reik, W. (2017). Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Reports*, *20*(5), 1215–1228. https://doi.org/https://doi.org/10.1016/j.celrep.2017.07.009

Regev A, L. E., Teichmann SA. (2017). Human cell atlas meeting participants. The human cell atlas. *Elife*. https://doi.org/10.7554/eLife.27041

Rodriguez-Terrones, G., D. (2018). A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nature Genetics*. https://doi.org/10.1038/s41588-017-0016-5

Stuart T, H. P., Butler A. (2019). Comprehensive integration of single-cell data. *Cell*. https://doi.org/10.1016/j.cell.2019.05.031